RADC-TR-65-314
Final Report

DISSEMINATION RESEARCH

Peter G. Ossorio

TECHNICAL REPORT NO. RADC-TR-65-314
December 1965

Information Processing Branch
Rome Air Development Center
Research and Technology Division
Air Force Systems Command
Griffiss Air Force Base, New York

FOREWORD

This final technical report has been prepared under Contract AF30 (602)-3432 by Peter G. Ossorio, Department of Psychology, University of Colorado, Boulder, Colorado. The work was performed under Project 4594, Task 459401 from July 1964 to July 1965.

The RADC Project Engineer was Mr. Robert N. Ruberti, EMIIH.

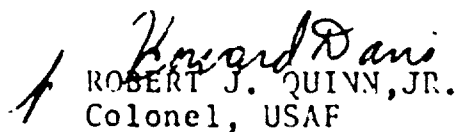This technical report has been reviewed and is approved.

Approved:

FRANK J. TOMAINI
Chief, Info Processing Branch
Intel and Info Processing Branch

WILLIAM H. HARRIS
MAJOR, U. S. AIR FORCE

Approved:

ROBERT J. QUINN, JR.
Colonel, USAF
Chief, Intel and Info Processing Div.

FOR THE COMMANDER:

IRVING J. GABELMAN
Chief, Advanced Studies Group

ii

ABSTRACT

A description is given of the initiation of an opera-
tional dissemination system based on the Classification
Space methodology developed previously. A set of three
evaluational studies, now in progress, is described.

A report is made of the empirical results of three
semantic studies designed to form the basis for complementing
the Classification Space by providing a capability for
indexing and dissemination in terms of the conceptual content
of documents or other textual units.

Recommendations for the further development of the
dissemination system and for the further application of
pragmatic methodology in linguistic data processing are made.

TABLE OF CONTENTS

EVALUATION

The results described in this report establish a completely new automatic method for organizing and disseminating technical information. This new method, designated as an "Attribute-Space", is a substantial departure from any previous procedure used in linguistic data processing. The outstanding feature of this method is the following; it is based on the premise that the principles by which technical people select and use information are complex and multidimensional and that perhaps, the reason why other indexing systems fail, is because they are only unidimensional.

The author states this premise on p.23. "For dissemination as well as for more complex IS&R operations it is essential to have an ordering of data (here, documents) which corresponds to the principles of selection employed by the users of the system. For real users, the principles which determine the acceptability and relative importance of documents received from an IS&R system are complex and multidimensional rather than simple or unidimensional. Indeed, the complexity of information requests is only somewhat less than the complexity of language itself. The level of success attributable to current indexing system appears to be achieved primarily by building one type of selective principle into the system and supplementing this with heuristic programming or its equivalent, together with the ingenuity and patience of the user."

As it is pointed out later in the report, the major principle that has always been used for classifying text has been to organize it in terms of different subject matter or different fields of knowledge only. Unfortunately, this approach has proven to be inadequate time and time again. This effort has demonstrated that information can also be organized in terms of "conceptual content". To organize or classify documents in terms of "conceptual content" is meant to organize in terms of the kind of information found in the document. Subject matter classification does not identify kinds of information but rather simply places documents within different domains of activity. An example of classification based on subject matter only would be the ability to identify a document which is about "radar". A classification system based on both subject matter and conceptual content, however, could differentiate between documents dealing with radar antennas, radar hardware or radar systems. The differences between these three is not with respect to subject or the domain of activity, but with respect to the different kinds of information involved. For example the kind of information involved in radar hardware deals with concepts such as weight, shape, size and substance while radar antennas deal with mathematical and time and space concepts as well as physical concepts.

Radar systems might be considered to deal with all of these plus theoretical concepts.

Accordingly, subject content of documents can be categorized or classified in terms of a new entity. The procedures for obtaining valid measurements of this entity are described in this report.

In addition this report gives a description of the initiation of an operational dissemination system based on the Classification Space methodology developed previously.

1.0 Introduction

Previous studies in linguistic data processing (Ossorio, 1964) have provided empirical evidence for the viability of a pragmatic methodology exemplified by the Classification Space technique for subject matter indexing. In the pragmatic approach an attempt is made to map into the LDP system significant aspects of the ways in which linguistic data enters into the activities of the users of that data. A significant advantage of a successful attempt of this sort is that the ordering of the data within the system corresponds to the principles which determine the acceptability and relative importance of the linguistic data to the users of the system. Thus, for example, data within a Classification Space is ordered in terms of subject matter relevance, and this reflects the fact that subject matter relevance relative to a topic of interest is one of the primary bases for acceptance and rejection of documents by users.

At the close of the experimental demonstration reported in RADC - TDR - 64 - 287, two directions for immediate further development were indicated. The first was to employ the Classification Space method in an operational setting in order to provide the basis for further development and meaningful evaluation. The initiation of such an effort is reported in Section 2.0. The

second line of development was to provide the experimental

foundation for a different form of data-ordering, specifically,

an ordering in accordance with the conceptual content of the

data. The initial empirical work for such an effort is reported

in Section 3.0.

2.0 Dissemination Studies

The initiation of a Classification-Space-based dissemina-

tion system at the Rome Air Development Center provided the op-

portunity for evaluation of the Classification Space method in

an operational setting. Briefly, the Dissemination Studies con-

sist of one constructive study and three evaluational studies.

In the constructive study a Classification Space was constructed

for a content domain determined by the range of interest of RADC

users. The three evaluational studies consist of systematic

procedures for assessing the effectiveness of the dissemination

selections for particular users. The constructive study is

described in Section 2.1, and the design of evaluational studies,

which are in progress at the present time, is reported in Section

2.2.

2.1 The RADC Classification Space

A survey of the subject matter interests of the expected

users of the RADC Dissemination System resulted in the identi-

fication of 75 fields of knowledge which collectively defined

the domain of interest to these users. The Classification Space study was designed to provide coverage of this 75-field domain. However, as in the earlier Classification Space study (RADC - TDR - 64 - 287) it proved impossible to obtain sufficient expert informants to provide coverage of the entire domain of interest. This practical limitation resulted in the restriction of the final Classification Space to a domain consisting of 49 of the 75 fields. These 49 fields are listed in Table 1.

A detailed description of the procedures for constructing a Classification Space was presented in an earlier report (RADC - TDR - 64 - 287) and will not be repeated here. Briefly, the procedure involves selecting a number of technical expressions from the literature of the fields of knowledge comprising the domain of interest and obtaining scaled judgments as to the degree of relevance of each of the technical expressions to each of the fields of knowledge. The judgments of relevance of the terms to a given field are made by informants who are professionally competent in that field and judgments for the several informants in a given field are averaged. The result is a two-dimensional data matrix reflecting the relevance of each term to each field. Here, the fields are treated as variables and are intercorrelated on the basis of the relevance data. The result, for K fields, is a KxK correlation matrix. The correlation

matrix is then factor analyzed and the result is a NxK factor matrix which can be interpreted as an N-dimensional Euclidean space in which is embedded a configuration of K vectors (corresponding to the K fields) extending from the origin of the space. The configuration is determined by the fact that, within the limits of the factor analytic approximation, the cosine of the angle between any two field vectors is equal to the numerical value of the correlation between the two fields. The configuration of vectors represents the collective scope of the K fields, and the reference axes of the space provide a systematic frame of reference for representing this content domain.

Point locations within the space may be assigned to linguistic units such as words, phrases, sentences, paragraphs, or documents. What is required is a quantitative estimate of the degree of relevance of the linguistic unit to each of the K fields (or some effective subset of these fields). These estimates can be interpreted as projections of the linguistic unit on the various field vectors. Then, since the projections of the field vectors on the reference axes of the space are given by the results of the factor analysis, it is possible to estimate the projection of the linguistic unit on the reference axes. When a metric is adopted for the space, the estimation of these latter projections is equivalent to assigning a set of

coordinates, hence a determinate location within the space, to the linguistic unit.

Because the data upon which the factored correlation matrix is based is relevance data, the space may be interpreted as a relevance space so that to assign a location to a linguistic unit in this space is to characterize that unit in respect to its degree of relevance to any actual or possible field of knowledge which is effectively represented as a vector within the space. Thus, the assignment of coordinates in the relevance space to a linguistic unit is equivalent to classifying it in terms of its subject-matter relevance within the content domain defined by the space. It is for this reason that such a space is designated as a "Classification Space".

Indexing a linguistic unit by obtaining relevance judgments by expert informants with respect to a substantial number of fields of knowledge would be an unwieldy way of processing text in an operational setting. The previous Classification Space studies showed that it was possible to approximate effectively expert judgments regarding the relevance of paragraph units by making use of the relevance coordinates of technical expressions appearing in the paragraph units. This was accomplished by using from four to six of the technical expressions which occurred in each paragraph and applying the Classification Formula in order to arrive at a location for the paragraph as a

whole. Thus, the operation of a Classification Space IS&R system on a fully automatic basis would depend on the availability of an already-indexed set of terms, or "system vocabulary", which was sufficiently large to provide the practical assurance that for the documents being processed at least four of the technical expressions occurring in the document would also be found in the system vocabulary.

The data matrix for the RADC Classification Space represented 49 fields and 1459 technical expressions. Three sets of technical expressions were selected independently:

(a) Following the standard Classification Space procedure developed previously twelve technical expressions were chosen randomly from six documents selected as belonging to the literature of a given field. This was done for each of the 75 fields, defining the RADC users' domain of interest. A total of 900 technical expressions was selected in this way.

(b) Each of the 75 fields was selected by one or more of the prospective users of the dissemination system. For each field, the user who selected that field was asked to generate six terms which he considered to be most distinctively representative of that field. The purpose of this procedure was to achieve at least a minimum assurance that the terminology characteristic of each field would be taken into account in the construction of the Classification Space. A total of 450 tech-

6

nical expressions was generated in this way.

(c) because the system vocabulary requires the same kind of relevance judgments as those on which a Classification Space is based, a portion of the Classification Space terms was selected with the intention of using the terms as the system vocabulary for the functioning system. Since abstracts from the Foreign Technology Division constitute the principal source of documents routinely available to the RADC users, the terms for the system vocabulary were taken from this source. The selection was made by determining a set of terms which would ensure that each of a set of 500 documents obtained successively from the Foreign Technology Division via existing channels of distribution would be represented in the system vocabulary by at least five technical expressions if the abstract itself contained that many. Under this criterion, it was established that a set of 587 terms would suffice and that for the last hundred of the 500 documents each additional document required an average of one additional term to the system vocabulary, so that the expectancy for any additional documents to be processed would be that at least four technical expressions in each document would be found in the system vocabulary.

Thus, a total of 1937 technical expressions were selected from the three sources. When the overlap of selections from the three sources was eliminated, the total was reduced to the final figure of 1459.

On the basis of the 49 x 1459 matrix of relevance judgments, the 49 fields were intercorrelated and factor analyzed. Twenty-nine orthogonal factors were extracted by means of the Maximum Likelihood method. These were rotated in accordance with the Varimax criterion. After rotation, sixteen of the 29 factors retained appreciable loadings by one or more variables (i.e., a minimum of .500). The sixteen factors accounted for 72 per cent of the total variance. In addition, five of the 49 fields were poorly represented in the factor space. Each of these five was added as a separate, independent reference axis, resulting in a 21-dimensional Classification Space.

A summary of the factor results is presented in Table 2. For each factor, the fields are listed in descending order of magnitude of loadings and only fields having loadings of .400 or higher are included.

The results of the analysis are highly interpretable, and no anomalous relationships are found. The configuration of fields in the 16-dimensional common factor space conforms to what would be expected on the basis of a general knowledge of the nature of the various fields. Of the sixteen factors, nine are associated with fields which have sufficiently high loadings to permit effective measurement. The remaining seven are marginal in this respect. Thus, indexing and measurement in this Classification Space will occur in less than optimal conditions.

8

Table 1
List of Fields for RADC C-Space

1. Adaptive Systems
2. Antennas
3. Applied Mathematics
4. Associative Processors
5. Audio Engineering
6. Ballistic Missile and Satellite Detection
7. Circuit Theory
8. Communication Theory
9. Computers
10. Computer Memories
11. Computers in Command and Control
12. Computer Software
13. Control Theory
14. Crystallography
15. Digital Circuitry
16. Digital Communications
17. Digital Storage Devices
18. Display Consoles
19. Document Storage and Retrieval
20. Electric Fields
21. Electroacoustics
22. Electromagnetic Fields
23. Electronic Data Processing
24. Electronic Recording Systems
25. Electro-optics
26. Feedback Control Systems
27. Field Theory
28. Logic
29. Logic Circuitry
30. Magnetic Fields
31. Maintainability
32. Microelectronics
33. Microwave Networks
34. Non-numeric Data Processing
35. Numerical Analysis
36. On-Line Processing
37. Parallel Computer Organization
38. Pattern Recognition
39. Phased Array Radar Systems
40. Probability and Statistics
41. Programming Languages
42 Reliability
43. Solid State Systems and Devices

Table 2

RADC Classification Space Analysis

Factor I   Digital Computer Data Processing

.853 Electronic Data Processing
.834 On-Line Processing
.829 Computers
.821 Programming Languages

.787 Document Storage & Retrieval
.783 Non-Numeric Data Processing
.735 Computer Software
.712 Parallel Computer Organization

.673 Associative Processors
.612 Computers in Command & Control

.583 Computer Memories
.575 Digital Storage Devices
.507 Electronic Recording Systems
.504 Logic Circuitry

.451 Pattern Recognition
.419 Digital Circuitry

Factor II   Audioelectronics

.842 Audio Engineering
.727 Electroacoustics

Factor III   Applied Mathematics

.882 Applied Mathematics
.854 Numerical Analysis
.300 Probability & Statistics
.748 Stochastic Processes

.494 Communication Theory
.476 Adaptive Systems
.463 Reliability
.411 Logic

Factor IV   Field Theory

.890 Field Theory
.842 Magnetic Fields
.817 Electric Fields
.771 Electromagnetic Fields

.617 Microwave Networks

Factor V   Microelectronics

.771 Microelectronics
.680 Solid State Systems & Devices

Factor VI   Detection Systems

.770 Phased Array Radar Systems
.712 Ballistic Missile & Satellite Detection
.687 Antennas
.591 Tracking & Prediction Theory

Factor VII   Communication Systems

.751 Digital Communications
.648 Wire Communications
.474 Telemetry
.420 Communication Theory

Factor VIII Digital Storage Devices

.687 Digital Storage Devices
.566 Computer Memories
.521 Digital Circuitry

Factor IX   Maintainability

.719 Maintainability
.717 Reliability

Factor X Electro-Optical Phenomena

    .656 Crystallography
    .585 Spectroscopy
    .413 Electro-Optics

Factor XI Adaptive Systems

    .572 Adaptive Systems
    .537 Pattern Recognition

Factor XII Control Systems

    .639 Feedback Control Systems
    .622 Control Theory

Factor XIII Display Consoles

    .650 Display Consoles
    .491 Electro-Optics

Factor XIV Telemetry

    .502 Telemetry

Factor XV Logic

    .572 Logic

Factor XVI Superconducting Circuits

    .658 Superconducting Circuits

"UNIQUE" FACTORS ADDED

Factor XVII	.751	Spectroscopy
Factor XVIII	.704	Crystallography
Factor XIX	.654	Maintainability
Factor XX	.566	Electro-Optics
Factor XXI	.549	Phased Array Radar Systems

2.2 Dissemination Evaluation Studies

Three evaluation studies are currently in progress, and no results are available at the present time. The procedures for these studies are described in Sections 2.2.2 to 2.2.4. As a preliminary to these, a description is given of the general characteristics of indexing and retrieval (or dissemination) in a Classification Space.

2.2.1 Classification Space Indexing and Retrieval

The paradigm case (and the simplest case) of indexing and retrieval is the following:

(a) Documents are indexed by being assigned a set of coordinates in the Classification Space by means of the Classification Formula and the system vocabulary.

(b) A retrieval request is interpreted by being assigned a set of coordinates in the Classification Space on the basis of the Classification Formula and the system vocabulary. Although the computations are the same for documents and retrieval requests, document coordinates are treated as point-locations for the document whereas the request coordinates are treated as the entry point of a request vector into the Classification Space. The differential implication of this procedure is that a document which is located close to the origin of the space presents no special problems, whereas a request which is indexed close to the origin is one which can-

not be responded to effectively by the system because essentially the entire content domain of the Classification Space is irrelevant to the request. The rationale and empirical confirmation of this differential implication is presented in RADC - TDR - 64 - 287.

(c) The distance between a document location and the request location is treated as an index of the degree of relevance of the document to the request. Thus sequential retrieval (or dissemination priority) is determined by the relative distances from the request to the available documents in the Classification Space. Those documents which are closest to the request are treated as being most relevant, hence are retrieved first or given highest dissemination priority. A vindication of this procedure is presented in RADC - TDR - 64 - 287.

The procedure described above involves a single point-location for the request and a selection criterion which is constant across all the dimensions of the Classification Space. More complex procedures are required in cases where (a) a request is assigned more than one location (i.e., a request of the form "something relating to X or Y or Z", or (b) the selection criterion is not constant across all the dimensions of the Classification Space (i.e., where instead of a spherical geodesics we have cylindrical, ovoid, or other forms, or (c) when the request is represented by a volume rather than

a point location. An operational dissemination system provides an appropriate setting for exploring the advantages and difficulties associated with these more elaborate procedures which are made possible by the geometric nature of the Classification Space.

2.2.2 Retrieval Efficiency as a Function of Document Source

In order to evaluate the efficiency of a retrieval or dissemination procedure, some criterion of effectiveness is required. The present studies employ two standard criteria, i.e., (a) the Relevance Ratio (RR), which is defined as the ratio of relevant items to the total number of items retrieved, and (b) the Selection Ratio (SR), which is defined as the ratio of the relevant items retrieved to the total number of relevant items available for retrieval. A more informative variation of the Relevance Ratio is obtained by having retrieved items ranked or rated as to their degree of relevance to a particular request and correlating this array with the rankings of the same items on the basis of their Classification Space distances from the request. This index is designated as the Relevance Correlation. Finally, both the Selection Ratio and the Relevance Ratio may be computed by assigning differential weights to each item based on the relevance ranking, so that the selection or non-selection of

16

the most relevant items contributes more to either the SR or the RR than the less relevant items do. The several measures described above are designated as Criterion Measures.

The first evaluational study is an investigation of the possible differential efficiency of dissemination depending on the source of the documents being processed. For this purpose, documents from three different sources are used. The three types of documents are ASTIA abstracts, FTD abstracts, and journal abstracts.

The following procedure is employed:

(1) Each user is assigned to a request location in the Classification Space on the basis of his expression of interest in particular subject matter areas.

(2) From an available machine-readable set of 1000 ASTIA abstracts, 400 FTD abstracts, and 100 journal abstracts. a random selection is made of 250 ASTIA abstracts and 300 FTD abstracts; all 100 journal abstracts are used.

(3) The 450 documents selected are processed by the dissemination system and ranked in their order of relevance to each user.

(4) For each user, the 20 most relevant ASTIA items, the 10 most relevant FTD items, and the two most relevant journal items are selected to form part of a test sample.

17

(5) From the 750 ASTIA items and 300 FTD items not selected previously, 80 of the former and 30 of the latter are selected at random; for each user, 38 journal abstracts are selected from the 98 which remain after the initial selection of the two most relevant items. The items selected in this stage are combined with the items selected in stage (4) to make up the total test sample.

(6) The total test sample is divided into two identically-stratified halves; the basic test procedure is performed once with the first half and replicated with the second half of the total test sample.

(7) In the basic test procedure, each user is given a set of 90 items. These consist of 10 relevant and 40 randomly selected ASTIA items, 5 relevant and 15 randomly selected FTD items, and 1 relevant and 19 randomly selected journal items. The user does the following: (a) classifies each item as "relevant" or "irrelevant" to his area of interest; (b) classifies the "relevant" items into three categories of degree of relevance ("most", "intermediate", "least"); (c) within each of the three categories he ranks the items in order of relevance; (d) he rates each item on a three-point scale of value or importance to him.

(8) On the basis of the information obtained from the user, the criterion measures are calculated. The Relevance Correlation is based only on the items judged to be relevant by the user.

2.2.3 Dissemination Efficiency as a Function of Text Conditions

The first dissemination study involves the use of abstracts as the documents processed by the system. One reason for this is the greater availability of abstracts in machine-readable form. Another is the fact that these abstracts, as a class of documents, are a major resource actually used by the RADC users. On the other hand, there is always an appreciable likelihood that a given abstract does not adequately represent the content from which it is abstracted. Too, there is some reason to expect that by virtue of the capability for indexing a document paragraph by paragraph , or even sentence by sentence, the Classification Space method may be particularly effective in indexing documents in full text rather than abstracts. The second dissemination study is a preliminary step in the direction of testing this expectancy. In it, the effectiveness of the dissemination procedure for full text as against the corresponding abstracts is investigated. Limitations on the availability of full text documents in machine-readable form requires that the experimental sample be limited

to 20 documents now classified as belonging in the field of
Information Processing. The following procedures are involved:

(1)   Three users from the RADC Information Processing Branch
serve as the experimental subjects. The Classification Space
user locations established for these users in the previous
study are used here.

(2)   The 20 documents and abstracts are processed by the dis-
semination system and ranked in order of relevance for each
user.

(3)   Each user makes the set of judgments needed for the cri-
terion measures. Abstracts are rated first, inasmuch as
the sequence effects involved in going from full text to the
corresponding abstract would be experimentally unmanageable.
Statistical comparisons between criterion measures for ab-
stracts as against full documents are made for each criterion
measure.

2.2.4   Dissemination Effectiveness as a Function of Request
        Interpretation

The difference between the simplest retrieval paradigm
(single point locations for user and spherical search volumes)
and more complex procedures was indicated above. The third
dissemination study involves a comparison of dissemination
effectiveness under that simplest condition as against one in

which the user request location is represented as a volume
in the Classification Space. At the same time, it involves
a comparison between request locations based on users' descrip-
tions of their areas of interest as against an inductive
method of determining those areas of interest. In this study
the parameter of induction vs description is not separated
from the parameter of point location vs volume location. An
effort to separate the two would be indicated if significant
differences were found. The following procedures are involved:

(1) Three or more (as available) RADC users serve as ex-
perimental subjects.

(?) A sample of 90 ASTIA, FTD, and journal items is selected
by the same procedures as were used in selecting the two test
samples for the first dissemination study.

(3) Criterion measures are the same as for the other two
dissemination studies.

(4) The data from both replications of the first dis-
semination study are used in order to plot into the Classi-
fication Space the following four sets of points for each
user: (a) the locations of the "most relevant" choices;
(b) the locations of items which were judged as being of
"intermediate" relevance; (c) the locations of the "least
relevant" items; and (d) the locations of the "most important"
items.

(5) One or more mathematical expressions are selected for describing the volume within which each of the four sets of points lies; these expressions satisfy the further condition that a determinate calculation of "the distance from an outside point to the user locations (to the surface of the volume)" is possible. The determination of efficient descriptions of this kind is one of the technical problems resulting from the increased flexibility in retrieval procedures made possible by the geometric character of the Classification Space.

(6) The third test sample of 90 items is processed by the dissemination system. Each item is categorized as "inside" or "outside" for each of the four volumes derived for each user. The distances from the "outside" items to the user volume are computed.

(7) Taking each volume for each judge separately, 12 "inside" and 18 "outside" items are selected at random from the third test sample.

(8) The 30 items so selected are presented to the user, who makes the judgments required for the criterion measures.

(9) For each of the criterion measures a comparison is made between the results obtained with the third test sample and the results obtained with the two test samples in the first dissemination study. In this comparison conflicting expect-

ations are involved. A drop in effectiveness may be expected by virtue of the unique variance, including error variance, of the original data. Significant decreases are the rule rather than the exception in cross-validational studies. On the other hand, an increase in effectiveness would be expected on the grounds that the selections for the third dissemination study are based on considerably more information than those for the first study. In the long run, for users with stable interests, it seems that the inductive approach would be practically certain to offer a significant improvement over the use of single descriptions. Thus, the present study might be regarded less as a test of whether the inductive approach is better than as an indicator of how much and how soon the superiority is shown.

3.0 Semantic Studies

For dissemination as well as for more complex IS&R operations it is essential to have an ordering of data (here, documents) which corresponds to the principles of selection employed by the users of the system. For real users, the principles which determine the acceptability and relative importance of documents received from an IS&R system are complex and multidimensional rather than simple or unidimensional. Indeed, the complexity of information requests is only somewhat less than the complexity of language itself. The level

of success attributable to current indexing system appears
to be achieved primarily by building one type of selective
principle into the system and supplementing this with
heuristic programming or its equivalent, together with the
ingenuity and patience of the user.

One major principle for selection of documents is
"subject matter relevance". This is a social psychological
principle which operates at the level of distinguishing
among the professional activities of groups of scientific
persons working in the same "field of knowledge". Under-
standably, "field of knowledge" units are molar rather than
molecular--they cannot reach the level of specificity of
some user needs. (For example, "the specific gravities of
chemical compounds of Type X" is too specific to qualify as
a "field of knowledge".) The Classification Space method
of subject matter indexing makes it possible to process
user requests of a more specific sort. For example, the
request for documents relevant to such topics as "the syn-
thesis of fat" and "vector analysis" were successfully
handled in an experimental study. (Cf RADC-TDR-64-287.)

Because the Classification Space method is a recent de-
velopment in linguistic data processing, little can be said at
the present time as to the limits of specificity of subject
matter that can be organized and identified in this way. No

doubt there are such limits. What is more important, in re-
gard to meeting user needs, is that subject matter relevance,
even in the multidimensional Classification Space format, is
only one principle of selection of information. It is to be
expected that the addition of other selective principles would
offer a superior basis for meeting user needs than is avail-
able in a system operating on the basis of the single princi-
ple of subject matter relevance.

A second major principle for selecting and evaluating in-
formation lies in the conceptual content of the information.
Roughly speaking, subject matter classification places a
document in relation to a domain of activity, whereas con-
ceptual content classification identifies the kind of infor-
mation contained in the document.

An example of the differential operation of these two
principles is the following: Given the ability to identify
documents relevant to "radar" as a subject matter, the User
might attempt to achieve a more discriminating request by
specifying "radar antennas", "radar hardware", or "radar sys-
tems". Any of these might or might not be effectively imple-
mented by a Classification Space in which "radar" was the
most specific field of knowledge incorporated into the indexing
space. A different way of achieving a more discriminating re-
quest would be to specify documents which (a) were relevant to

"radar" and (b) the content of which dealt with concepts such as weight, shape, size, and substance rather than (c) mathematical concepts or (d) space and time concepts. This selection would probably come close to the selection which would be appropriate to "radar hardware". And a combination of (b) and (c) would give something more like the selection which would be appropriate to "radar antennas", whereas (c) would bias the selection in the direction of the theoretical aspects of radar. There are parallels of this kind between a subject matter selection and a subject matter plus conceptual content selection, but it is clear that in general no one-to-one correspondence is to be expected. Equally, selection by means of conceptual content should not be regarded as merely a way of approximating very specific subject matter distinctions. Rather, it is a second basic principle in terms of which users make information selections and evaluate information selections. This conclusion is based initially on general psychological principles and is substantiated by the results of interviewing a number of users, including RADC personnel.

Thus, the present Semantic Study is an attempt to provide an indexing system for conceptual content which could subsequently be used in implementing the conceptual content selection of documents in an operational IS&R system. The technical approach is basically analogous to that for the Classification Space,

26

though with some important modifications. Such an indexing system for conceptual content is designated as an "Attribute Space".

The following is a description of the basic procedures for constructing an Attribute Space. These are the procedures for performing factor analysis and factor measurement, and they are described in greater detail in standard textbooks on factor analysis.

(1) A conceptual domain is delimited by a set of logical predicates formulated in a sentential format (for example, "X is microscopically small"). The range, or content, of this conceptual domain is, within the limitations of the precision of measurement, just the range of the particular set of predicates used. No effort is made to sample an independently described content domain except at the very general level of "the physical world". (A selection which emphasized biological, psychological, social, or other content domains would also be possible.) The primary effort is to employ a set of predicates which would effectively represent the gross conceptual dimensions of the physical world. It is these dimensions which would correspond to the coordinate axes of the Attribute Space. The selection of predicates is based on previous work in this regard (Ossorio, 1961). Both a priori considerations and empirical consensus are reflected in the selection. Lists of predicates are given in Appendices A, B, and C.

(2) A domain of application (of the predicates) is delimited

by a set of "Objects" to which the predicates are applied.

The function of the set of Objects is to provide a significant

frame of reference for assessing empirically the similarities,

or redundancies among the various predicates. That is, two

predicates are treated as being similar to the extent that their

application across the set of Objects is similar. Thus, the

Objects represent a comprehensive sampling of objects, events,

and actions. The list of Objects, broadly categorized, is given

in Appendix D.

(3) Given a set of predicates and a set of Objects, judges are

asked to compare each predicate with each Object and to rate

the degree to which the predicate is applicable to the Object.

Ratings are made on a 9-point rating scale. The instructions

to the judges are presented in Appendix E.

(4) The complete set of ratings of each predicate with respect

to each Object is designated as one "replication". (The Semantic

Study made use of average data from ten replications.) The

averaged data is designated as the "Protocol" or "protocol data"

for the Semantic Study.

(5) Using the protocol data, the predicates are intercorrelated

and the matrix of intercorrelations is factor analyzed. On the

basis of the Classification Space Analysis and previous analyses

of semantic data, either the Minimum Residual method or the Maximum Likelihood method of factor extraction in conjunction with the Varimax criterion for rotation of the reference axes are considered to be among the appropriate factor analytic procedures for analyzing the data. In the present studies, the Maximum Likelihood method was used.

(6) The result of the factor analysis is a Euclidean space defined by N orthogonal coordinate axes. In this space the predicates are represented as vectors of standard length fanning out from the origin of the coordinate system. The configuration of vectors has the property that the cosine of the angle between any two vectors is, within the limits of the factor analytic approximation, equal to the correlation coefficient for the corresponding predicates.

(7) The appraisal of the conceptual dimension represented by a given reference axis in the coordinate system is made in the light of an examination of those predicate vectors which have the highest projections on that axis, since those are the predicates the conceptual content of which is most strongly associated with the conceptual dimension represented by the coordinate axis.

(8) The ratings of a given Object with respect to a given predicate may be interpreted as the projection of an "Object vector" on the predicate vector. If a given coordinate axis

29

has one or more predicate vectors sufficiently strongly associated

with it, the ratings of an Object with respect to the predicate

vectors can be used to estimate the projection of the Object

on the reference axis. In this way the Object is given a set

of coordinates with respect to the orthogonal reference axes

which represent conceptual dimensions. The computation of

coordinates for a given Object in this way constitutes the

indexing of that Object in the Attribute Space--it is classified

in terms of its conceptual content.

(9) The classification of documents in an Attribute Space poses

the same kind of problem as the classification in a Classification

Space. That is, to classify documents on the basis of a limited

system vocabulary consisting of terms already classified in

the space. It is this feature which makes possible the fully

automatic indexing of documents. In the Classification Space,

a practically effective solution has been provided by the

Classification Formula. For the Attribute Space, some attention

will have to be devoted to the examination of the various possible

formulae for computing document coordinates as a function of

the coordinates of terms which appear in the document.

Two kinds of complication must be introduced in connection

with the Attribute Space as contrasted with the Classification

Space. The first has to do with the question of technical vs.

non-technical concepts, and the second relates to different kinds

of predicate expressions.

The Semantic Study was designed to result in an Attribute Space for a domain of ordinary language predicates. There are two reasons for this. First, there is substantial informal evidence that the differences between technical and non-technical language is not that the former covers an extensive domain of unique conceptual content, but rather, that in it we find conceptual content organized in different "packages". Thus, we might expect that an Attribute Space constructed from a broad set of non-technical predicates would be adequate to represent all or most of the conceptual content of technical expressions.

This likelihood makes possible the solution of a practical difficulty, i.e., that the amount of data (hence the man-hour participation) required for an Attribute Space is substantially greater than would be required for a Classification Space of comparable complexity. Most likely, it would not be feasible to construct an Attribute Space coordinate with the RADC Classification Space if the construction of the former required data obtained entirely from technical personnel in the way that a Classification Space does. The conceptual point of an expected high degree of overlap between technical and non-technical conceptual content domains and the practical point of the availability of technically competent participants are both involved in the planning of the Semantic Study.

Given an Attribute Space covering a conceptual domain
relevant to scientific and technical fields, the use of the
Attribute Space for dissemination or retrieval would require
the indexing of a set of technical terms in the space. These
terms, the "system vocabulary", would be the same as the system
vocabulary for the Classification Space in order to be used
effectively in conjunction with the latter. It is when the same
terms are indexed in both spaces that the system can most
effectively implement dissemination or retrieval requests which
combine subject matter principles of selection with conceptual
content principles of selection. The indexing of a system
vocabulary does require the rating of each term by technically
competent personnel. The number of terms rated and the number
of judges used for each rating may vary within wide limits, and
certainly, a prototype Attribute Space already constructed for
a non-technical domain can be put to use in a technical domain
with considerably less data than would be required to construct
an Attribute Space for a technical domain.

The second complication has to do with the fact that there
are several broad classes of predicate expressions. For the
present study, three major classes were selected. These are
designated as (a) simple predicates, (b) functors, and (c)
categories. The following are examples of descriptions of these
three kinds, respectively: (a) "X is large", (b) "The size of

X is important", and (c) "X is primarily something mental".

The point of distinguishing among the three types is illustrated by relating each of the three examples to a single Object, "A Mistake": (a) It is clear that no simple predicate expression such as "X is large" or "X is small" is particularly applicable to the Object "A Mistake". This is because mistakes come in all sizes, and so a mistake as such is neither small or middle-sized or large. (b) Nonetheless, o.e of the most significant features of a mistake is whether it is a large mistake or a small one. By and large, large mistakes are serious and small mistakes are not. Thus the functor "The size of X is important" is highly relevant to the Object "A Mistake". Whenever a type of Object is characterized by significant variability along some dimension of analysis it will be the functor associated with that dimension rather than simple predicates associated with particular values on the dimension which will be most applicable to an unspecified Object of that type. (c) In both simple predicative and functorial characterizations, the focus is an analytic one--it is aspects or dimensions of the Object that are referred to in this way. In categorial characterizations, on the other hand, the focus is on the classification of the Object in its entirety--the Object as a whole is grouped together with other, similar Objects. Thus, it is a mistake as such, rather than some aspect of it, that "is primarily mental" rather than, e.g., physical or chemical.

Similarly, a cat as such is an animal, and its being an animal is different from its being furry or having claws.

Because each of the three types of predicate expressions contributes separately to the descriptive characterization of Objects, the Semantic Study was divided into three sub-studies in which the predicates were, respectively, functors, categories, and properties. An Attribute Space constructed on such a basis may be expected to provide a more adequate representation of the range of conceptual content found in scientific and technical literature.

Functors, categories, and predicates were analyzed separately. These analyses are reported below.

3.1 The Functor Analysis

The functor analysis was based on ten replications of a 79x191 data matrix which represented ratings of 191 Objects with respect to 79 functor expressions. The latter are listed in Appendix B. The 191 Objects, used for all three semantic studies, are listed in Appendix D. The factor results are summarized in Table 3. Thirty-six factors, accounting for 88% of the total variance, were extracted. After rotation, 26 factors, accounting for 84% of the total variance retained significant projections by one or more of the functors. The high proportion of common variance to total variance is reflected in the fact that among the 79 variables in the analysis only six showed communalities lower than .80, with the lowest being .696.

34

An examination of the factor results shows no anomalous combinations among the variables most highly associated with any given factor. In a number of cases the variables associated with a given factor correspond closely to the prior informal grouping of the functors indicated by the alphabetical designations in Appendix B. The mass-weight-density factor is an example.

Among the 26 factors, 16 were characterized by a sufficiently high association with at least one variable to make factor measurement feasible (a maximum loading of .70 or higher was the criterion used to assess feasibility). Thus, although not all the conceptual content dimensions which were identified as a result of the analysis can be effectively measured on the basis of the present results, even 16 dimensions would provide a very substantial degree of differentiation and classification for indexing and retrieval purposes.

In the study, three of the functor expressions were used twice in the collection of data. The members of each pair were treated as separate variables in the factor analysis, so that the latter was a 79-variable analysis even though only 76 distinct functor expressions were used. The purpose of this procedure was to provide an informal check on the reliability of the data in terms of factor loadings, as contrasted with the more common correlational indices of reliability. A second purpose was to check on the degree to which introducing some high

Table 3

Functor Space Analysis

Factor 1   Identity of Individuals

.913   The identity of X is important
.909   The identity of X is important
.821   You have to distinguish each X from every other X
.739   The history of X is important
.525   The origin of X is important
.555   The variability of X's is important
.487   The analysis of X is important
.454   The number of X's is important
.422   The internal characteristics of X are important

Factor 2   Cost

.989   The man-hour cost of X is important
.895   The man-hour cost of X is important
.814   The time cost of X is important
.789   The monetary cost of X is important
.389   The means-ends characteristics of X are important

Factor 3   Emission and Radiation Characteristics

.879   The emission characteristics of X are important
.849   The radiation characteristics of X are important
.474   The energy required for X is an important consideration

Factor 4   Space-time Dynamics

.892   The velocity of X is important
.862   The rapidity of X is important
.739   The movement of X is important
.698   The dynamic properties of X are important
.600   The temporal progression of X is important
.582   The dynamic balance of X is important
.543   The energy required for X is an important consideration
.501   The temporal sequence of X's is important
.486   The flow of X is important
.430   The rate of change of X is important
.400   The control of X is important

Factor 5   Weight-mass-density Characteristics

  .873   The weight of X is important
  .766   The mass of X is important
  .434   The density of X is important

Factor 6   The Factual Implication

  .843   The consequences of X are important
  .837   The immediate circumstances associated with X are
        important
  .772   The immediate effect of X is important
  .711   The implications of X are important
  .662   The outcome of X is important
  .629   It's important to avoid X
  .601   The long-term effects of X are important
  .472   The momentary state of X is important
  .444   The origin of X is important
  .424   The beginning of X is particularly important
  .414   The proof of X is important

Factor 7   Pragmatic Validity

  .792   The range of error for X is important
  .787   The rigorousness of X is important
  .786   The validity of X is important
  .782   The precision of X is important
  .715   The test of X is important
  .632   The proof of X is important
  .579   The efficiency of X is important
  .557   The amount of skill required for X is important
  .523   The means-ends characteristics of X are important
  .516   The success of X is important

Factor 8   Observable Form

  .783   The form of X is important
  .632   The shape of X is important
  .393   The outside of X is important

Factor 9   Constituent Composition

  .781   The chemical composition of X is important
  .516   The density of X is important
  .380   The physical constituents of X are important
  .342   The internal characteristics of X are important

Factor 10

.779 The observable characteristics of X are important

Factor 11 Spatial-structural Characteristics

.776 The boundary of X is important
.769 The distance to X is important
.749 The distance to X is important
.744 The spatial extent of X is important
.545 The size of X is one of its distinctive features
.414 The physical constituents of X are important

Factor 12

.751 The numerical range of X is important
.344 The number of subdivisions of X is important

Factor 13

.750 The later portions of X are particularly important
.696 The end of X is important
.572 The beginning of X is particularly important
.519 The outcome of X is important
.427 The amount of skill required for X is important
.409 The temporal progression of X is important
.402 The success of X is important

Factor 14

.735 The usual condition of X is important
.724 The maintenance of X is important
.581 It's important to know what state X is in
.523 The capacity of X is important
.497 The outcome of X is important
.423 The velocity of X is important
.378 The control of X is important

Factor 15

.725 The amount of X is important

Factor 16 Structural Characteristics

.715 The sub-structures of X are important
.649 The part-whole characteristics of X are important

.613 The number of subdivisions of X is important
.450 The internal characteristics of X are important
.336 The analysis of X is important

Factor 17

.677 The flow of X is important

Factor 18 Productivity

.640 The productiveness of X is important
.604 The output of X is important

Factor 19 Duration

.741 The duration of X is important
.623 The temporal span of X is important
.386 The rate of change of X is important

Factor 20

.591 The number of X's is important
.276 The size of X is one of its distinctive features

Factor 21

.538 It's important to avoid X

Factor 22

.524 The means-ends characteristics of X are important

Factor 23

.456 The capacity of X is important

Factor 24

.454 The control of X is important

Factor 25

.445 Access to X is important

Factor 26

.438 The long-term effects of X are important

correlations into the correlation matrix would affect the factor structure. In all three cases both members of the pair showed highly similar factor loadings. In two of the three cases the pair of variables represented the highest two loadings on one factor. For the third pair this was not the case, although the pair did show substantial loadings on one factor. Thus, the reliability ch ck was highly satisfactory and the check on the stability of the orientation of the reference axes in the presence of local ed artifacts was inconclusive. (Very nearly identical results were obtained from the double use of three simple predicate expressions in the analysis of properties.)

3.2 The Category Analysis

The analysis of categories was based on ten replications of a 49×15? data matrix representing the ratings of 191 Objects w. ... respect to 49 category descriptions. The latter are listed in Appendix A.

The factor results are summarized in Table 4. Twenty-four factors accounting for 83% of the total variance were extracted in accordance with the Maximum Likelihood method. Nineteen factors, accounting for 80% of the total variance, were retained after rotation. The Varimax criterion was used. Among the 49 variables, 13 showed communalities of less than .80, the lowest being .634. Among the 19 factors, 12 were characterized by sufficiently high loadings (.749 or higher) to make factor measurement feasible.

40

Table 4

Category Space Analysis

Factor 1   Electromagnetic Phenomena

    .972   X is primarily electromagnetic
    .921   X is primarily electrical
    .910   X is primarily magnetic
    .543   X is primarily energy-transforming

Factor 2   Biological Phenomena

    .947   X is primarily biological
    .886   X is primarily physiological
    .835   X is primarily organic

Factor 3   Conceptual vs physical

    .882   X is primarily imaginary
    .829   X is primarily mental
    .816   X is primarily hypothetical
    .805   X is primarily conceptual
    .624   X is primarily speculative
   -.617   X is primarily tangible
   -.610   X is primarily physical
    .534   X is primarily tentative
    .491   X is primarily logical
    .459   X is primarily evaluative
    .434   X is primarily affirmative
    .433   X is primarily linguistic

Factor 4   Mathematical Phenomena

    .868   X is primarily numerical
    .665   X is primarily statistical

Factor 5   Geometric Phenomena

    .834   X is primarily geometric
    .777   X is primarily spatial
    .569   X is primarily structural
    .365   X is physical

Factor 6   Instantiation

    .824 X is primarily observational
    .447 X is primarily illustrative

Factor 7   Temporal Phenomena

    .819 X is primarily sequential
    .756 X is primarily temporal
    .743 X is primarily periodic
    .504 X is primarily kinetic
    .455 X is primarily transitional

Factor 8   Experimental Phenomena

    .790 X is primarily experimental
    .785 X is primarily empirical
    .565 X is primarily procedural
    .478 X is primarily speculative

Factor 9   Intelligence

    .785 X is primarily information-transforming
    .720 X is primarily linguistic
    .673 X is primarily illustrative
    .594 X is primarily logical
    .489 X is primarily affirmative
    .465 X is primarily evaluative
    .451 X is primarily conceptual

Factor 10   Mechanical devices

    .764 X is primarily mechanical
    .670 X is primarily technological
    .568 X is primarily structural

Factor 11   Conventions-norms

    .756 X is primarily conventional
    .625 X is primarily normative
    .545 X is primarily social

Factor 12

    .749 X is primarily recreational
    .706 X is primarily artistic

Factor 13 <u>Chemistry</u>

    .692 X is primarily chemical
    .476 X is primarily energy-transforming
    .362 X is primarily causal

Factor 14 <u>Productivity</u>

    .688 X is primarily productive
    .264 X is primarily technological
    .254 X is primarily affirmative

Factor 15

    .659 X is primarily final
    .322 X is primarily causal
    .311 X is primarily procedural

Factor 16

    .615 X is primarily transitional

Factor 17

    .531 X is primarily self-correcting

Factor 18

    .494 X is primarily tentative

Two other factors were marginal in this respect (maximum loadings of .692 and .688).

As in the functor analysis, no anomalous combinations of categories were found. Although no attempt was made to group the category expressions prior to the analysis, the groupings exhibited by the factor pattern show a high degree of conceptual unity and thus are relatively easy to interpret as dimensions of conceptual content with respect to which documents could be ordered.

3.3 The Analysis of Properties

The analysis of properties was based on ten replications of a 101x191 data matrix representing the ratings of 191 Objects with respect to 98 simple predicate expressions. The latter are listed in Appendix C.

The factor results are summarized in Table 5. Thirty-nine factors, accounting for 85% of the total variance were extracted. Of these, 31 factors, accounting for 80% of the total variance, were retained after totation. Among the 31 factors, 16 were characterized by sufficiently high loadings (.752 or higher) to make factor measurement feasible, and five additional factors were marginal in this respect (maximum loadings of .657 to.698).

No anomalous combinations of proper ies were found among the high-loading properties associated with any given factor.

Table 5

Property Space Analysis

Factor 1   Negative Evaluation

    .935 X is bad
    .919 If only X could be gotten rid of
    .630 Something should be done about X

Factor 2   Electromagnetic Phenomena

    .920 X is electromagnetic
    .878 X transmits energy
    .863 X is magnetic
    .793 X radiates energy
    .721 X changes in a matter of microseconds
    .702 X conducts electricity
    .678 X contains a lot of energy
    .640 X receives energy
    .628 X requires a lot of energy
    .513 X has a large spatial range
    .450 X changes in seconds
    .440 X has a finite range
    .409 X is very rapid

Factor 3   Lack of Reality

    .898 X is imaginary
    .844 X is unreal
    .660 X is intrinsically unobservable
    .651 X is subjective
  -.473 X is observable
    .463 X has no known limit or end
  -.427 X is a very clearcut sort of thing

Factor 4   Demand Characteristics

    .874 X requires special attention
    .845 X requires special attention
    .832 X requires constant attention
    .630 X has to be controlled at all times
    .610 X is complicated
    .545 Every X is a special case
    .528 X has to be taken through successive stages
  -.504 X is simple and undifferentiated
    .469 X has complex constituents
    445 X combines alot of other things into one
    .416 X has to be done one step at a time
    .401 X is important in its own right

45

Factor 5   <u>Decision Monitoring</u>

.844   X is correct
.825   X should retain its relative position
.783   X has a small range of uncertainty
.696   X is valid
.615   X is parametric
.535   X is axromatic
.426   X has to work just right or it's no good
.418   X progresses in an orderly fashion
.412   X has a finite range
.390   There is a standard form for X

Factor 6   <u>Observable Individuality</u>

.839   X has a definite shape
.828   X has definite boundaries
.774   X is discrete
.769   X has several colors
.733   X is highly structural
.714   X has a regular boundary
.672   X is complete in itself
.651   X should remain in the same condition
.619   We can recognize X when we encounter it
597   X should retain its relative position
.588   X is heavy
.571   X is irregularly shaped
.5_9   X is observable
.557   X is a very clearcut sort of thing
-.515   X is an intermittent process
.485   X changes over a period of years
.432   X is dense
.423   Most X's are pretty much alike
.415   X has complex constituents
.415   X is normally in constant balance
.397   There is a standard form for X

Factor 7   <u>Macrocosmic Characteristics</u>

.832   X is astronomically large
.637   X is far away
.429   X is large
.422   X has no known beginning
.420   X has a large spatial range

Factor 8   Origination

    .817 X occurs only under specific conditions
    .811 X occurs only under specific conditions
    .546 X has a known cause
    .456 X has to be generated in a particular way

Factor 9   Active Phenomena

    .800 X changes over a period of days
    .746 X shifts from one state to another
    .712 X is changed by its own action
    .689 X shifts from one form to another
    .421 The hidden qualities of X are the important ones

Factor 10
    .802 X is microscopic
    .329 X is hard to distinguish from its surroundings

Factor 11

    .795 X is all or none

Factor 12   Goal Focus

    .784 The important thing is X no matter how you arrive at it
    .638 X is good
    .418 X is important in its own right

Factor 13   Identifiability of particular instances

    .775 Instances of X can be recognized immediately
    .447 We can recognize X when we encounter it
    .414 X is observable

Factor 14   Means Focus

    .774 X is a very effective means
    .762 X is a means to an end
    .453 X has to work just right or it's no good
    .438 There is a standard form for X
    .421 X by itself replaces alot of things

Factor 15   Momentum

    .760   X is very rapid
    .539   X is hard to stop once it is started
    .385   X changes in seconds

Factor 16

    .752   X is recursive
    .379   X is an intermittent process
    .311   X is part of a definite sequence

Factor 17   Process Focus

    .698   X has a beginning, middle, and end
    .642   X develops in a regular way
    .641   X has to be taken through successive steps
    .629   X progresses in orderly fashion
    .575   X is gradual
    .555   X develops in a regular way
    .545   X has to be done one step at a time
    .439   X is part of a definite sequence

Factor 18

    .696   X has a characteristic color
    .282   X is simple and undifferentiated

Factor 19

    .681   X is dense
    .633   X is heavy

Factor 20   Part Focus

    .675   X is part of a definite structure
    .534   X is part of a larger aggregate
    .384   X should retain its relative position

Factor 21

    .657   X is hard to distinguish from its surroundings
.   .242   X develops in a regular way

Factor 22

    .632   X is non-linear
    .328   X changes in a matter of microseconds
    .306   X changes in seconds

Factor 23

    .621 X is linear
    .304 X has a regular boundary

Factor 24

    .602 X is topological
    .394 X has a large spatial range

Factor 25 Bounded vs Unbounded

    .543 X has no known limit or end
 -.313 X has a beginning, middle, and end

Factor 26

    .487 X requires occasional attention

Factor 27

    .481 X has to be generated in a particular way

Factor 28

    .437 X shifts from one form to another'

Factor 29 Sources vs media

    .442 X conducts electricity
 -.333 X radiates energy

Factor 30

    .413 Every X is a special case
    .292 X is important in its own right

Factor 31

    .422 X is normally in constant balance

Factor 32 Standardization

    .389 Most X's are pretty much alike
    .359 There is a standard form for X

The factor structure showed a substantial relationship to the prior grouping of the predicate expressions although nothing approaching a one-to-one correspondence was found.

The inclusion of three "doublets" produced results highly similar to the results obtained with the three doublets in the functor analysis. The results are inconclusive, but they do suggest that although introducing a doublet has a tendency to affect the final orientation of at least some of the reference axes with respect to the configuration of vectors, that configuration is sufficiently strongly patterned so that this influence is not always decisive and does not result in anomalous factors.

3.4 An Attribute Space for Indexing

As indicated above, the requirements for an Attribute Space which would serve as a complement to a Classification Space are simply (a) that appropriate judgments be obtained as data for a factor analysis, (b) that the factors resulting from the factor analysis be measurable, and (c) that the terms in the system vocabulary of the Classification Space also be measured and indexed within the Attribute Space. On the basis of the present study, only the last of these conditions has yet to be met. The three analyses reported above provide a total of 44 measurable factors.

Although there is a substantial overlap in the descriptions
of factors in the three analyses, a serious loss in indexing
capability would result from eliminating factors whenever two
or more factors in the different analyses appear to represent
the same conceptual content dimension. The distinctions among
categories, functors, and properties remains a significant one
even when the same conceptual content dimension is involved.
This was illustrated previously in the difference between "size"
properties and the "size" functor in relation to the Object
"a mistake". Thus, the most effective use of the results of
the present study appears to be as an Attribute Space which is
composed of three distinct sub-spaces.

4.0 Summary Discussion

The initiation of a dissemination system operating on
Classification Space principles has been described and the ini-
tial results of semantic studies designed to provide a basis
for an operational conceptual content indexing capability have
been reported.

Because the dissemination system is proto-typical and does
not replace a pre-existing system, none of the parameters of the
functioning system has the status of a "given" for either ser-
vice procedures or evaluation. Consequently, a good deal of

latitude is left open in both respects. Equally, no quick and simple evaluation procedure which is both conceptually coherent and empirically convincing is possible. There is not, in fact, widespread agreement or conviction as to what would qualify as a criterion for evaluating an information system. The degree of consensus which exists at the present time centers around the two measures designated in Section 2 as the "Relevance Ratio" and the "Selection Ratio". A disadvantage of such indices as they are now used is that they put a very heavy burden on the user to say what he wants, what he would have chosen if..., and what it was he actually used. Although reliance on such judgments is likely to remain an irreducible aspect of evaluation methodology, judgments of this kind will be maximally informative if they are obtained against a background of some systematic analysis of the needs of users and their patterns of usage of information.

In practice, therefore, it would seem that an effective evaluation procedure would be a cyclical process in which the initiation of system procedures was followed by some empirical evaluation which formed the basis for initiation of further changes followed by further evaluation, etc. In any functioning system, some balance is eventually achieved between the

adaptation of system resources to the criterion decisions made by users and the adaptation of users to the system as they become better able to make criterion judgments or engage in search or request activities which enable them to exploit the resources of the system. Only when stable patterns of usage emerge is it possible to perform sensitive and informative evaluation studies by varying parameters of the user-system interaction. To arrive at such a point with respect to the present dissemination system requires not only a continuing study of users and their needs, but also sampling decisions (i.e., for sources of documents or for additions to the system vocabulary), data gathering (i.e., of the kind in Section 2), and programming (i.e., implementing the more complex retrieval procedures described in Section 2) which might be required to implement the changes suggested by previous evaluation outcomes.

Two aspects of Classification Space technology appear to have sufficient long run significance to warrant immediate attention whether or not they meet any immediate practical requirements. The first is the possibility that subject matter discrimination can be substantially increased by the construction of "minature" Classification Spaces in which the scope of the content domain is greatly limited and subject matter descriptions at a more specific level than "field of knowledge" descriptions

are used as variables. An associated problem, assuming that miniature spaces are feasible, is that of tying in the miniature spaces to more comprehensive spaces for purposes of indexing or retrieval.

The second technical problem is the development of efficient approximative methods for altering or expanding a Classification Space indexing structure without involving a major empirical effort. Preliminary effort on this problem at the present time would be of significant value for later study of the problem of updating a Classification Space and for developing and evaluating procedures for coping with the latter problem.

The results of the semantic studies demonstrate a degree of technical adequacy which provides a firm empirical basis for expecting that an Attribute Space based on these results would make a significant contribution to the data processing capability inherent in the Classification Space method. Although many of the dimensions of conceptual content identified in these studies cannot be measured effectively at present, many of them can be. Since the problem of how to interpret the conceptual dimensions appears to be essentially nonexistent, there is reason to expect that those which are measureable will have a direct use and that measurement can be developed for those dimensions which so far cannot be

measured, because a knowledge of what these conceptual dimensions are makes it possible to sample them more adequately than in the present studies.

The addition of conceptual content indexing to the present dissemination system would involve the following: (a) the collection of data (semantic ratings) from informants who are competent in the technical and scientific fields relevant to the documents processed by the system, (b) the construction of an Attribute Space by methods basically similar to those used in constructing a Classification Space, (c) the programming of a set of indexing and retrieval operations for the Attribute Space, (d) the programming of Boolean operations relating subject matter relevance criteria to conceptual content criteria for implementing user requests, (e) the development of effective formats for expressing or formulating information requests in a system having the multiple and flexible capabilities provided by the combination of Classification Space and Attribute Space resources, and (f) some evaluation of the effectiveness of the Attribute Space as such and as an increment to the Classification Space capability.

As indicated in Section 3.0 the pragmatic approach to linguistic data processing involves the strategy of developing discriminative classification capability not by making finer

distinctions, through elaborate statistical techniques, within

a single kind of classification, but rather by using simple

methods to project linguistic data onto multiple frames of

reference which serve as models for the conceptual structures

of users which determine how they select and use linguistic

data. One of the methodological principles involved in this

strategy is that refining discrimination within a homogeneous

frame of reference (i.e., a single dimension) contributes addi-

tively to overall discriminative power whereas the combination

of disparate frames of reference contributes multiplicatively.

With respect to the multidimensional frames of reference

mapped into a Classification Space and an Attribute Space there

seems little question that these are among the primary bases

for processing linguistic data. Moreover, the problems in-

volved in using the two in combination appear to be primarily

practical and technical problems, Certain other frames of

reference, such as those provided by part-whole relationships

and means-ends relationships appear to be sufficiently salient

to warrant immediate preliminary empirical study. However, as

the number of actual or proposed frames of reference increases,

two questions also become increasingly important. These are

(a) "Why these frames rather than others?", and (b) "How do

they go together?".

Thus, there is a definite need at the present time for a comprehensive and relatively detailed formulation of the pragmatic conceptualization of language and of the methodological implications of this conceptualization for automatic linguistic data processing. A successful formulation of this kind would involve the description of a comprehensive program for research and development centering around answers to the questions of which frames and how they go together.

References

1. Ossorio, Peter G.
   Classification Space Analysis (RADC-TDR-64-287)
   Rome, N.Y.: RADC Systems and Technology Division, 1964

2. Ossorio, Peter G.
   Meanings in Ordinary Language
   Unpublished doctoral dissertation, UCLA, 1961

Appendix A

List of Categories

1. X is primarily biological
2. X is primarily tangible
3. X is primarily tentative
4. X is primarily speculative
5. X is primarily observational
6. X is primarily procedural
7. X is primarily causal
8. X is primarily transitional
9. X is primarily relational
10. X is primarily conceptual
11. X is primarily physical
12. X is primarily logical
13. X is primarily temporal
14. X is primarily spatial
15. X is primarily chemical
16. X is primarily social
17. X is primarily statistical
18. X is primarily empirical
19. X is primarily numerical
20. X is primarily experimental
21. X is primarily structural
22. X is primarily imaginary
23. X is primarily magnetic
24. X is primarily linguistic
25. X is primarily technological
26. X is primarily electrical
27. X is primarily final
28. X is primarily information-transforming
29. X is primarily electromagnetic
30. X is primarily periodic
31. X is primarily mathematical
32. X is primarily affirmative
33. X is primarily energy-transforming
34. X is primarily evaluative
35. X is primarily conventional
36. X is primarily hypothetical
37. X is primarily illustrative
38. X is primarily normative
39. X is primarily organic
40. X is primarily mechanical
41. X is primarily physiological

42. X is primarily geometric
43. X is primarily kinetic
44. X is primarily mental
45. X is primarily sequential
46. X is primarily productive
47. X is primarily recreational
48. X is primarily self-correcting
49. X is primarily artistic

Appendix B

List of Functors

Group

    1. It's important to avoid X

A   2. The control of X is important
    3. It's important to keep X within certain limits

B   4. The dynamic balance of X is important
    5. The dynamic properties of X are important

C   6. The weight of X is important
    7. The mass of X is important
    8. The density of X is important

D   9. The flow of X is important
    10. The movement of X is important
    11. The emission characteristics of X are important
    12. The radiation characteristics of X are important

E   13. The amount of X is important
    14. The number of X's is important

F   15. The velocity of X is important
    16. The rate of change of X is important
    17. The rapidity of X is important

G   18. The numerical range of X is important
    19. The variability of X's is important
    20. The observable characteristics of X are important

H   21. The usual condition of X is important
    22. It's important to know what state X is in

I   23. The static properties of X are important
    24. The form of X is important
    25. The shape of X is important

J   26. The energy required for X is an important consideration
    27. The efficiency of X is important
    28. The time cost of X is important
    29. The monetary cost of X is important
    30. The man-hour cost of X is important

31. The size of X is one of its distinctive features

K 32. You have to distinguish each X from every other X
33. The identity of X is important

L 34. The analysis of X is important
35. The amount of skill required for X is important
36. The precision of X is important
37. The range of error for X is important

M 38. The temporal progression of X is important
39. The temporal sequence of X's is important
40. The temporal span of X is important
41. The duration of X is important

N 42. The boundary of X is important
43. The spatial extent of X is important
44. The distance to X is important

O 45. The cause of X is important
46. The beginning of X is particularly important
47. The history of X is important
48. The origin of X is important

P 49. The later portions of X are particularly important
50. The end of X is important

Q 51. The consequences of X are important
52. The implications of X are important
53. The long-term effects of X are important
54. The immediate effects of X are important
55. The outcome of X is important

R 56. The immediate circumstances associated with X are important
57. The momentary state of X is important

S 58. The physical constituents of X are important
59. The chemical composition of X is important
60. The internal characteristics of X are important
61. The substructures of X are important
62. The microscopic structure of X is important

T   63. The part-whole characteristics of X are important
     64. The means-ends characteristics of X are important
     65. The productiveness of X is important
     66. The output of X is important

U   67. The rigorousness of X is important
     68. The proof of X is important
     69. The validity of X is important
     70. The success of X is important

     71. The number of subdivisions of X is important

     72. The test of X is important

     73. The capacity of X is important

     74. The access to X is important

     75. The outside of X is important

     76. The maintenance of X is important

Appendix C

List of Properties

1. X is large
2. X is microscopic
3. X is astronomically large

4. X is precise
5. X has a small range of uncertainty

6. X has definite boundaries
7. X has a regular boundary

8. X is hard to distinguish from its surroundings
9. X is a very clearcut sort of thing

10. X is far away
11. X has a large spatial range
12. X has a finite range

13. X receives energy
14. X transmits energy
15. X requires alot of energy
16. X contains alot of energy

17. X is electromagnetic
18. X radiates energy
19. X is magnetic
20. X conducts electricity

21. X is observable
22. We can recognize X when we encounter it
23. Instances of X can be recognized immediately
24. X is intrinsically unobservable
25. The hidden qualities of X are the important ones

26. X is unreal
27. X is imaginary
28. X is subjective
29. X is valid
30. X is correct

31. X is dense
32. X is heavy

33. X is complete in itself
34. X is changed by its own action

35. X is linear
36. X is non-linear
37. X is parametric
38. X is topological
39. X is axiomatic

40. X by itself replaces alot of things
41. X combines alot of things into one

42. X requires constant attention
43. X requires occasional attention
44. X requires only routine attention
45. X requires special attention

46. X has several colors
47. X has a characteristic color

48. X changes in microseconds
49. X changes in seconds
50. X changes over a period of days
51. X changes over a period of years

52. X has to be taken through successive steps
52. X is hard to stop once it is started
53. X has to be controlled at all times
54. X has to be done one step at a time

55. X has a known cause
56. X has no known beginning
57. X has a beginning, middle and end
58   X has no known limit or end

59. X is a means to an end
60. X is a very effective means
61. X is important in its own right
62. The important thing is X, no matter how you arrive at it
63. X has to be generated in a particular way
64. X occurs only under specific conditions

65. X develops in a regular way
66. X develops slowly
67. X progresses in an orderly fashion
68. X is part of a definite sequence

69. X is an intermittent process
70. X is part of an irregular sequence
71. X is recursive

72. X has to work just right or it's no good
73. X is normally in constant balance
74. X shifts from one form to another
75 X shifts from one state to another
76. X should remain in the same condition
77. X should retain its relative position

78. something should be done about X
79. If only X could be gotten rid of
80. X is bad
81. X is good

82. X is simple and undifferentiated
83. X has a definite shape
84. X is irregularly shaped
85. X is highly structured
86. X is complicated

87. X has complex constituents
88. X has simple constituents
89. X is part of a definite structure
90. X is part of a larger aggregate

91. X is continuous
92. X is discrete
93. X is all-or-none
94. X is gradual

95. There is a standard form for X
96. Every X is a special case
97. Most X's are pretty much alike

98. X is very rapid

Appendix D

List of Objects

I   <u>Invalidity</u>

 1. breaking a rule
 2. a false promise
 3. a sales pitch
 4. an erroneous proof
 5. an accident
 6. a mistake
 7. a dream

II   <u>Criteria</u>

 8. a definition
 9. a calculation
 10. a measurement
 11. an experiment
 12. a custom
 13. putting it to a vote
 14. a referee
 15. a textbook
 16. flipping a coin
 17. remembering something
 18. seeing it right there

III   <u>Pathology</u>

 19. an illness
 20. a stalled automobile
 21. a slow wristwatch
 22. dying
 23. a fit of coughing
 24. an earthquake
 25. an explosion
 26. a yawn
 27. an argument

IV   <u>Therapy</u>

 28. flushing a radiator
 29. a hospital
 30. mending a fence
 31. tuning a piano
 32. optimization

33. spring cleaning
34. curing an illness

V Contests

35. a chess game
36. a lawsuit
37. a hand-to-hand battle
38. broken-field running
39. keeping up with the joneses

VI Assertions--communications

40. giving a lecture
41. a radio broadcast
42. giving directions to someone
43. describing something
44. praying
45. persuading someone
46. a press release

VII Decomposition

47. cutting meat
48. grinding ore
49. taking a clock apart
50. analyzing an argument
51. decomposition

VIII Tools

52. a pair of pliers
53. a hand drill
54. a microscope
55. a blowtorch
56. a hose
57. a lock

IX Biological

58. a man
59. a tree
60. blood
61. a sweetheart
62. a moth
63. a virus
64. a seed

X Costs

65. buying something
66. paying a fine
67. a traffic ticket
68. being drafted
69. making a down payment
70. a gas bill

XI Creativity--discovery

71. a hunch
72. an inspiration
73. discovering something
74. exploring
75. inventing something
76. wondering about something
77. making something

XII Construction

78. building an airplane
79. moulding clay
80. hammering a nail into a plank
81. an assembly line
82. making a round hole

XIII Production

83. fertilizing the crop
84. a full tank of gasoline
85. a quantum of energy
86. a computer program
87. an atomic pile
88. being at bat
89. getting the answer
90. rotating crops

XIV Mechanisms

91. a clock
92. an IBM computer
93. a gas meter
94. the solar system
95. a television set
96. a guided missile
97. a train

XV Artifacts

98. a radar antenna
99. a cradle
100. a dollar bill
101. a pair of snowshoes
102. a lens
103. a bear trap
104. a dart
105. a calendar
106. a high-voltage wire
107. a workbench
108. a milk bottle

XVI Structures

109. a claw
110. a building
111. a lattice
112. a crescendo
113. a wire
114. a piece of lace
115. a bubble
116. a blob
117. an arrow
118. a slab
119. a sheet
120. a box

XVII Natural objects

121. a river
122. a cloud
123. a shadow
124. a boulder
125. the sun
126. a valley
127. a flame
128. an island
129. the ocean

XVIII Aggregates--quantities

130. a combination
131. a beginners' class
132. a nation
133. a square dance
134. adding more of the same
135. a collection
136. a heap of stones
137. a pound of meat
138. a ton of metal
139. a pile of wood

XIX Fruition

140. harvesting wheat
141. splitting the profits
142. declaring a dividend
143. a glass of beer
144. a hearty meal

XX Representations

145. a pencil sketch
146. a portrait
147. a map
148. a blueprint
149. a theory
150. a photograph
151. an explanation
152. diagnosis

XXI Miscellaneous

153. a table of random numbers
154. a railroad schedule
155. radio waves
156. a beacon
157. adding a pinch of salt
158. a bright light
159. sound
160. music
161. asking somebody
162. an infinite set
163. dripping water
164. the ticking of a clock
165. a pleasant mood
166. excitement
167. a novel
168. a chance encounter
169. penetrating a barrier
170. crossing over a river
171. going around a mountain
172. a log in the road
173. comparing two samples
174. calibrating a compass
175. a candidate
176. a criminal
177. a refugee
178. running a business
179. a novel
180. a gas bill

181. a foggy night
182. having your luck run out
183. imitating someone
184. taking something for granted
185. good health
186. a pinch of salt
187. the weather
188. having a strong suspicion
189. buying a lottery ticket
190. the direct wire to Moscow
191. an exceptional case

Appendix E

Semantic Study Instructions

ORIENTATION

This is a data-gathering procedure in which you will be
asked to make judgments based on your knowledge of certain
common objects, actions, events, or situations. (For con-
venience, the word "object" will be used here to refer to
either an object, an action, an event, or a situation.)

For each object, you will be given a set of descriptive
statements, and your task is to decide to what extent the
description applies to the object. For example, the object
might be "an illness" and the description might be "the con-
sequences of X are important". Here, you would consider "an
illness" to be the "X" in the statement and you would judge
to what extent the consequences of an illness are important.

You would express your judgment by making a checkmark on
a scale like this:

an illness

```
|___|___|___|___|___|___|___|___|___|
  0   1   2   3   4   5   6   7   8
```
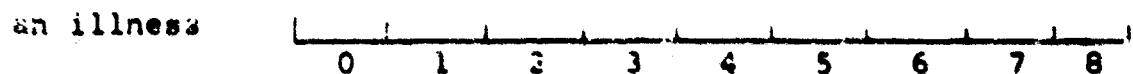
In general, the greater the degree to which the description
applies to the object the higher should be the number that you
check on the scale. Keeping this general principle in mind, use
the following as a guide in making your ratings:

1. Check "0" if the description <u>doesn't</u> <u>apply</u> <u>at</u> <u>all</u> to the ob-
    ject. For example, it may be definitely false to describe
    the object that way, or it may not make any sense at all to
    describe the object that way.

2. Check <u>either</u> "1" or "2" if the description <u>applies</u> <u>only</u> <u>to</u>
    <u>a</u> <u>minimal</u> <u>degree</u>, but you wouldn't want to say that it
    doesn't apply at all. For example, it may make sense but
    be far-fetched to describe the object that way, or the
    description may apply but only in a very qualified or
    restricted sense. If you are inclined to say "Yes, you
    could say that, but . . . .", then "1" or "2" is an ap-
    propriate rating.

3. Check <u>either</u> "3" or "4" if the description <u>does</u> <u>apply</u> <u>but</u>
    <u>is</u> <u>relatively</u> <u>uninformative</u>. For example, it may refer to
    a trivial or incidental feature of the object, or it may
    apply to only a minority of the specific instances covered
    by the "object" expression.

4. Check <u>either</u> "5" or "6" if the description <u>definitely</u> <u>ap-</u>
    <u>plies</u> <u>and</u> <u>is</u> <u>informative</u>. For example, the description may
    refer to a significant feature of the object, or it may
    represent what is normally to be expected of the object, or
    it may refer to a characteristic which, though not a usual
   one for the object, is significant when it is present.

74

5. Check either "7" or "8" in the most important or significant cases of a description applying to an object. For example, the description may refer to a defining characteristic or a necessary characteristic of the object, or to one of the most crucial or outstanding features of an object, or to a characteristic which would be absent only in very special circumstances.

In deciding between "1" and "2", "3" or "4", "5" or "6", "7" or "8", use the general rule that the greater the degree to which the description applies to the object, the higher the number that you should check on the scale.

On each page of your booklet you will find one description at the top of the page and below, twelve scales with the object given alongside. Take each of the objects in turn and relate it to the description, making your checkmark on the corresponding scale each time.

IMPORTANT:

1. Rate each item in turn. Do not skip any.

2. Make your check mark in the middle of the scale sections, not on the divisions:

```
      This                      Not This
  |___|___|_✓_|___|___|___|___|___|_✓_|___|
    0   1   2   3   4   5   6   7   8
```
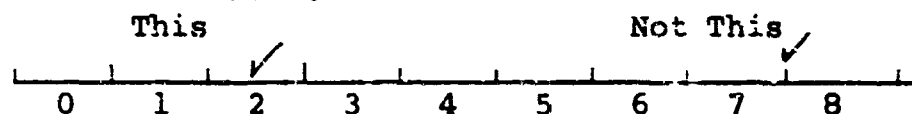
3. Make each judgment independently. Do not try to remember how you rated other objects or descriptions. Take each page in order. Do not look back and forth in your booklet.

75

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*	2a. REPORT SECURITY CLASSIFICATION
Dr. Peter G. Ossorio	UNCLASSIFIED
Department of Psychology	**2b GROUP**
University of Colorado, Boulder, Colorado	

3. REPORT TITLE

Dissemination Research

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

July 64 - July 65

5. AUTHOR(S) *(Last name, first name, initial)*

Ossorio, Peter G.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
December 1965	90	2
8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)	
AF30(602)-3432	n/a	
b. PROJECT NO.		
4594		
c. Task No.	9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*	
459401	RADC-TR-65-314	
d.		

10. AVAILABILITY/LIMITATION NOTICES

Distribution of this document is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	RADC (EMIIH), GAFB, N.Y.

13 ABSTRACT

A description is given of the initiation of an operational dissemination system based on the Classification Space methodology developed previously. A set of three evaluational studies, now in progress, is described.

A report is made of the empirical results of three semantic studies designed to form the basis for complementing the Classification Space by providing a capability for indexing and dissemination in terms of the conceptual content of documents or other textual units.

Recommendations for the further development of the dissemination system and for the further application of pragmatic methodology in linguistic data processing are made.

DD FORM **1473**
1 JAN 54